REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-
0563

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE 8/8/2000

3. REPORT TY... ...COVERED Final Report 3/1/99 - 2/29/00

4. TITLE AND SUBTITLE
DURIP: Object Recognition in Cluttered Scenes Using Compressed Data

5. FUNDING NUMBERS
Contract FQ 8671-9900942
Grant F49620-99-1-0122
Task 3484/US

6. AUTHOR(S)
Octavia I. Camps

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Pennsylvania State University
Dept. of Electrical Engineering
University Park PA 16802

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NM
801 N Randolph St Room 732
Arlington, VA 22203-1977

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFOSR/NM

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release, distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Robust, reliable methods for automatic target/object detection, recognition, classification and identification are key technology areas for meeting the U.S. Air Force requirements for defense operations in warfare, as well as in peacekeeping and humanitarian role situations.
However, the recognition of general three-dimensional objects in cluttered scenes remains a challenging problem. In particular, the design of a good representation suitable to model large numbers of generic objects that is also efficient and robust to occlusion and segmentation problems, while minimizing probabilities of false alarm and misdetection, has been an stumbling block in achieving success.
To address this problem we have developed a new representation and theoretical models for object recognition based on appearance-based parts (ABPs) and relationships (ABRs), obtained from collections of images compressed using the Wavelet transform. This representation will allow us to design a recognition system that overcomes the problems mentioned above and that can work directly on compressed data, during both the training and the recognition stages, making it both time and memory efficient.

20001030 094

14. SUBJECT TERMS

Object recognition, image segmentation, image compression

15. NUMBER OF PAGES 25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

DURIP: Object Recognition in Cluttered Scenes Using Compressed Data

Final Report

Dr. Jon Sjogren
Program Manager
AFOSR/NM
Air Force Office of Scientific Research
801 N. Randolph Street Room 732
Arlington, VA 22203
Tel: (703) 696 – 5999
email: *jon.sjogren@afsor.af.mil*

by

Octavia I. Camps
Department of Electrical Engineering
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802
Tel: (814) 863–1267
Fax: (814) 865–7065
email: *camps@whale.ee.psu.edu*

Abstract

The PI requested funds for the acquisition of a 3D sensor for capturing range data and for high performance computing workstations to renovate Penn State's Robust Purposive Vision Laboratory. This equipment is supporting research towards the development of both theoretical tools and specific algorithms for robust object recognition using compressed data. This research was initiated by the PI during her 1997 summer research visit to Eglin Air Force Base and it is the subject of a proposal to be submitted to AFOSR in the near future.

Robust, reliable methods for automatic target/object detection, recognition, classification and identification are key technology areas for meeting the U.S. Air Force requirements for defense operations in warfare, as well as in peacekeeping and humanitarian role situations. However, *the recognition of general three-dimensional objects in cluttered scenes remains a challenging problem.* In particular, the design of a good representation suitable to model large numbers of generic objects that is also efficient and robust to occlusion and segmentation problems, while minimizing probabilities of false alarm and misdetection, has been an stumbling block in achieving success.

To address this problem we have developed a new representation and theoretical models for object recognition based on *appearance-based parts* (ABPs) and *relationships* (ABRs), obtained from collections of images compressed using the Wavelet transform. This representation will allow us to design a recognition system that overcomes the problems mentioned above *and* that can work *directly* on compressed data, during both the training and the recognition stages, making it both time and memory efficient.

Contents

Part I

Project Description

1 Introduction

Robust, reliable methods for automatic target/object detection, recognition, classification and identification are key technology areas for meeting the U.S. Air Force requirements for defense operations in warfare, as well as in peacekeeping and humanitarian role situations. However, the recognition of general three-dimensional objects from 2D images of cluttered scenes remains a challenging problem. In particular, the design of an efficient representation suitable to model large numbers of generic objects that is also *efficient* and *robust* to occlusion, texture segmentation, and data uncertainty so that *minimum probabilities of false alarm and misdetection* can be obtained, has been an stumbling block in achieving success.

One of the major difficulties in recognizing three dimensional objects from 2D images, is that an object appearance changes significantly depending on the point of view it is observed from. Common approaches to overcome this problem are to use viewer-centered representations to describe the objects in terms of their appearance, or to use object-centered representations together with image invariants.

Viewer-centered representations can be as structured as features grouped into relational models within aspect views [4, 6, 8], or as loose as simply collections of model images [18, 26]. A major limitation of this type of approach is that it usually requires isolating the object of interest from the background. Thus, viewer-centered representations are difficult to use in the presence of occlusion and image clutter. Furthermore, at present, it is not clear how they could be used efficiently when the object libraries are large.

Approaches using object-centered representations such as part decomposition [3, 21, 34], have the potential to cope with both occlusion and large object databases. However, the definition of parts for generic objects and their image extraction remains a difficult problem [14].

To overcome the shortcomings of previous approaches, we have developed a new object representation and recognition methodology. In particular, we sought a representation capable of handling large number of free form objects in cluttered scenes and to develop a recognition engine capable of achieving acceptable performance in the presence of textured clutter, occlusion and data uncertainty, thus eliminating the need for *ad hoc* heuristics. The new representation is based upon the use of probabilistic models suggested by training data that has been compressed using a Wavelets decomposition. These models are being validated and will be tested under controlled experimentation. Since the system will work directly with compressed data, it will save time and memory. Furthermore, since the compression technique to be used is based on the Wavelet transform, the system will make use of frequency information made explicit by this transformation. Finally, the rigorous mathematical modeling will allow us to formalize the problem of minimizing the probabilities of false alarm and misdetection in the presence of uncertainty.

Funds provided by this grant were used to purchase specialized equipment required to facilitate the above tasks. In particular, sensors to capture image and range data and several computer workstations to efficiently segment the data were purchased. Data consisting of registered color and range images obtained with these sensors will be shortly available to computer vision researchers through the world wide web.

2 Objectives

The specific objectives of this research effort can be summarized as follows:

- **Development of an efficient new object representation to model large numbers of free form objects that is robust to both occlusion and segmentation problems.**

We propose to describe objects using a hierarchical structure of parametric eigenspaces representing sets of *appearance-based parts* (ABPs) and relations of ABPs – *appearance-based relations* (ABRs). We define ABPs in terms of closed regions segmented from *compressed* images using a texture segmentation algorithm based on the Minimum Description Length (MDL) principle, developed by the PI during her summer research visit to Eglin AFB. Since this new representation is learned from segmented images, it is able to represent free form objects and it is robust to segmentation problems. Since it is based on regions rather than on global properties, it is robust to occlusion. Since it is obtained from compressed data, it is both time and memory efficient. Finally, it will be able to handle large object databases, due to its hierarchical nature and memory efficiency.

- **Theoretical development of mathematical models for robust recognition from compressed uncertain 2D image data.**

Starting from appropriate noise models for the data, we are deriving robust probabilistic models to integrate prior information and compressed data to achieve *optimal performance.* In particular, we are deriving models for the *discriminatory power* of ABPs and ABRs and statistically validate them through experiments carried out using both simulated and real data, adhering to a rigorous experimental protocol. These models will be used to estimate confidence levels and accuracy of hypotheses made regarding object identity and pose, given an image observation. Modeling of the uncertainty will be used to attain *minimum probability of false alarms and misdetections* without resorting to the use of *ad hoc* parameters. Finally, the performance of the algorithms will be characterized in terms of efficiency as well as in terms of error metrics such as scene complexity, amount of clutter, compression rate, probability of misdetection, probability of false alarm, and probability of pose error.

In order to achieve these goals two important steps had to be addressed: we needed 1) an efficient, and reliable image segmentation algorithm capable of working with textured images and 2) to capture large amounts of realistic data to estimate the model parameters and to validate the proposed approach. Funds from this grant were used to purchase equipment to facilitate these tasks:

1. A Vivid 700 3D scanner from Minolta was purchased to acquire range data as well as registered

color textured images of isolated objects and cluttered scenes to simulate "ladar" images in our laboratory. In particular, this scanner provides autozoom as well as 8 steps of zoom to scan objects as close as 0.5m and as far as 3.5m from the sensor. A Pentium system was also purchased to be used as a host for the 3D scanner while a second one replaced an old Gateway with a 486 processor to control all our image acquisition equipment (camera, pan and tilt head, rotating stage).

2. As part of the proposed research we are conducting hundreds of thousands of experiments to model noise statistics and to characterize the performance of the proposed techniques. Funds from this grant were used to update the aging computing facilities in our laboratory that were severely hindering our research progress due to their lack of computational power.

3 Background

3.1 Model Representation for Object Recognition

The representation of models is critical to the problem of 3D object recognition and pose estimation. A good sample of papers on this area (up to 1993) can be found in [19]. More recently, two workshops addressing the specific issue of object representation for recognition [15, 29] have evidenced that the problem of designing a *"good"* representation remains largely unsolved.

Most 3D object representations for recognition purposes proposed until now can be classified into one of three major categories: primitive-, physics-, and appearance-based representations.

Primitive-based representations rely on geometric models of objects [2, 9, 13, 17]. In order to cope with occlusion, decomposition of the object into parts is often used. Binford [3], for example, proposes a 3D part definition based on function. In this approach, a *divide and conquer* method is utilized to decompose complex objects into a structural representation. Another example can be found in [34], where Zerroug and Medioni employ a high-level, volumetric part-based approach where a hierarchical extraction process applies generalized cylinders to group compound objects from boundaries, surface patches, and volumetric parts. However, although there is general consensus on the fact that part decomposition can help overcoming occlusion, there is no agreement on what a part should be. Furthermore, reliable extraction of parts from 2D image data remains a difficult problem.

Physics-based representations typically model shape as a mechanical system subject to forces reflecting material properties as well as smoothness and image constraints. These methods have been successfully used for modeling complex objects whose shape may vary over time. Examples of this approach are the works of Metaxas [24] and Pentland and Sclaroff [28, 33]. Metaxas proposed a deformable model by integrating mathematical methods from geometry, physics and mechanics. In particular, Lagrangian mechanics were used to convert the geometric parameters of the solid primitive, the deformation parameters, and the six degrees of freedom of rigid-body motion, into generalized coordinates or dynamic degrees of freedom. Pentland and Sclaroff, on the other hand, used a finite element method, where the eigenvectors of the finite element model of the shape were employed to formulate the physical model. However, these methods are better suited for 3D object recognition from 3D data or 2D object recognition from 2D data.

The appearance of a 3D object in a 2D image depends on its shape, its reflectance properties, its pose in the scene, and the sensor and illumination characteristics. During her doctoral work, the PI developed the image prediction system PREMIO [6, 7] to simulate image segmentations under varying conditions (viewing position, illumination, and image processing parameters). The simulations obtained with the system were combined into probability models of segmentation errors that were used by a Bayesian-based recognition system. The main limitations of this system are that it requires a CAD model of the objects and that the simulations are not as realistic as they should. Costa and Shapiro [8] and Pope and Lowe [30] have addressed these problems by learning segmentations from real images. However, their methods require to train the system for the different aspects of the objects making it unclear whether or not their systems would scale up when the database of 3D models becomes large.

Murase and Nayar [26] proposed a representation for the learning, recognition, and pose estimation of rigid objects using a large set of images, obtained by varying pose and illumination, stored as parametric manifolds in an eigenspace. Object translation and scalingwere taken care by normalizing the image size using the bounding box of the object. Mundy et al [25] presented an experimental comparison between Murase and Nayar's method (SLAM) and two geometric model-based recognition methods described in [32] (Lewis) and [35] (Morse). This study concluded that the two approaches complement each other. Appearance models have the advantage that they do not require formal models to describe objects while geometric approaches rely on formal models to derive pose invariant properties. The major drawbacks of SLAM are that it is very sensitive to segmentation, in particular occlusion, that it does not lend itself well to object categorization, and that incidental variations in appearance such as texture or surface albedo must be modeled as separate objects. The major drawback of the geometric approach is that it is not robust to minor variations of the hard constraints imposed on the image geometry.

Recently, the appearance-based approach has received increasing attention [23, 22]. In particular, there has been a significant effort devoted to try to overcome the problems caused by occlusion and background clutter. In spite of these efforts, no satisfactory solution has been found to handle occlusion *without sacrificing scaling*. In [23] for example, a robust method to compute the coefficients to project an image into the parametric eigenspace was presented. This method extracts the coefficients by considering subsets of image points with a hypothesis-and-test paradigm and selecting the best hypothesis by using the MDL principle. As a result, the coefficients are robust to image outliers and in particular to occlusion. However, a major problem with this technique is that it cannot handle object translation and scaling. This is because this approach works only if the dimensions of the training and testing images are equal, and the pixel locations of the object do not change at recognition time. Unfortunately, occlusion has a direct impact on the object bounding box preventing the use of image size normalization in this case. In [22] Krumm proposed to handle occlusion by using small neighborhoods as features. Although this technique can handle object translation, it also suffers from scaling – i.e. it assumes that the object size in the image is the same at recognition and training time.

It should be noted that the representation proposed in this research is also related to the appearance-based representation proposed by Murase and Nayar. However, our approach significantly improves the representation robustness to segmentation problems and occlusion without compromising scaling or the ability to handle large databases of objects. This is accomplished by using local rather than global appearances, and by using the proposed representation as the main building block of a theoretical probabilistic model inspired on previous work by the PI. Further-

more, the proposed approach works with compressed data, improving storage and computation efficiency, and it will introduce a hierarchical representation suitable to group together "similar" parts into categories, allowing several objects to share parts.

3.2 Image Compression using Wavelets

Image compression is concerned with minimizing the number of bits required to represent images. Traditional applications of image compression are storage and transmission of images. It is only recently, that attention has been dedicated to its use for the development of fast algorithms that work directly on compressed data. The obvious advantages of using compressed data are that decompression of the data is avoided and that the algorithms process a reduced amount of data. This is particularly important for appearance-based approaches since they require a very large set of images to represent the models to be recognized. A less obvious advantage is that most compression techniques are based on transformations into the frequency domain that make frequency information, such as texture, more explicit.

A simple and yet powerful compression technique is based on using a pyramid image representation. From an original image, a coarse approximation is derived, for example by using a lowpass filter and a down sampler. The coarse image and the predicted error (the difference between the original image and the up sampled and filtered coarse image) can then be compressed. Reconstruction is then accomplished by adding the coarse image and the predicted error. This process can then be iterated over an over, forming a pyramid. One way of building this pyramid is to use a Wavelet decomposition. The Wavelet transform has the advantages that it provides good localization both in frequency and space domain and that its window size changes with the frequency content of the image. Furthermore, the Wavelet transform provides orientation sensitive information at various resolutions, and thus it is specially well suited for texture segmentation. Thus, we chose this compression technique as the basis for our segmentation algorithm. The pyramid obtained this way is then segmented using a multi-band multi-resolution segmentation technique based on the minimum description length.

To implement the Wavelet transform we use the same set of Quadrature Mirror Filters (QMF) that have been successfully used by Franques and coworkers [10]. The octave-band tree split is then obtained by splitting the lower half of the spectrum into two equal bands in the horizontal and vertical directions, at each level of the tree, as shown in Figure 1. The image is symmetrically extended before filtering, to reduce distortions due to boundary effects. This process generates at each level a coarse approximate of the input image and three orientation selective detail images, which are very important for texture segmentation, as shown in the next section.

4 Current Results and Future Work

In this research we proposed a new 3D object representation and a probabilistic model for the recognition of free form 3D objects from 2D images of cluttered scenes, suitable for semi-compressed data. The proposed research is being approached in two stages. First, the problem of designing a *good* representation, robust to both occlusion and segmentation problems that is also capable of

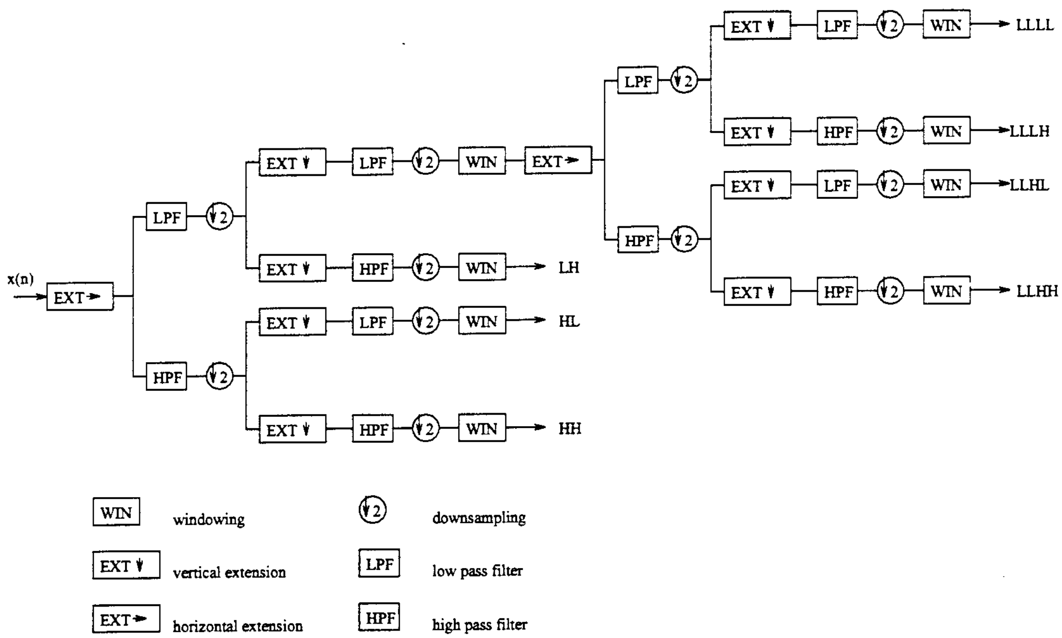


Figure 1: Decomposition of an image using a two-level dyadic tree.

modeling a large set of objects was addressed. The new representation, described in the following subsection, uses *appearance-based parts* combining properties from previous part- and appearance-based representations, but without many of their shortcomings. In particular, 1) it is learned from training images; 2) it describes objects in terms of local properties; 3) it is compact, and 4) it is obtained by working directly on compressed data.

In the second stage, we are in the process of deriving a robust probabilistic model for the *discriminatory power* of ABPs and groups of ABPs. This model will be used to 1) organize the object database in a hierarchical manner to handle the storage of large numbers of objects, and 2) formulate the problem of robustly identifying and locating objects in compressed images in a Bayesian framework, thus minimizing the probabilities of false alarms and misdetections.

4.1 Object Representation

Parts from Images

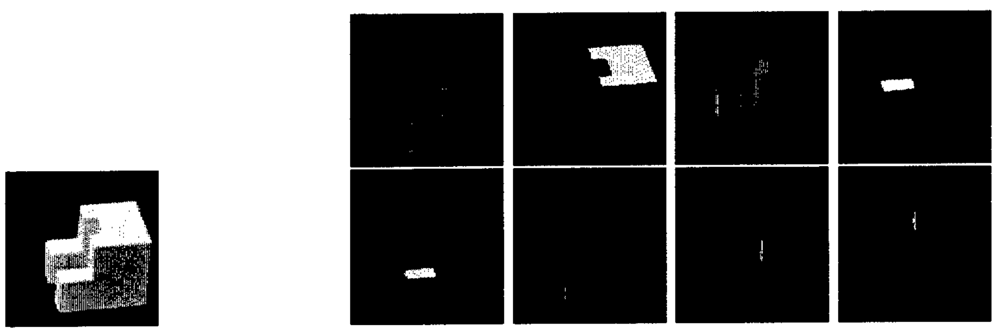
It is commonly accepted that complex objects can be decomposed into simple parts, and that part decomposition can help to overcome occlusion – i.e. a partially occluded object may be recognized if *enough parts* of it are recognized. However, there is no much agreement on how to define what a *part* is. Several definitions have been proposed in the past, including operational definitions (parts are what a part detector finds), view based definitions (parts are defined by local image properties), and geometric definitions (parts are defined by 3D events) [14].

We believe that, in order to be robust to segmentation problems, the definition of a part must take into account the segmentation algorithms that will be used to extract them from the images. In particular, we believe that a part definition should be used in the same way at the learning *and*

the recognition stages. Thus, we propose the following part definition:

> **Definition: Parts** are closed, non-overlapping image regions that optimally partition the image in a minimum description length (MDL) sense.

We have chosen an MDL based definition for the following reasons: 1) The MDL principle has a strong theoretical grounding: it is based on Occam's Razor, the principle which says that one should prefer the simpler of two theories explaining some data, everything else being equal. For MDL, Occam's Razor is applied in a coding sense, by fitting models to the given data, encoding the model parameters and the data using these models, and selecting the model that results in the smallest code length. This approach exploits the tradeoff existing between the complexity of a model and how well it fits the data. 2) Using MDL does not require arbitrary parameters, and thus parts can be extracted in a consistent manner. 3) The MDL objective function is formulated such that statistics are tested inside the regions and such that the resulting regions have homogeneous intensity (color or texture) properties. 4) Finally, efficient algorithms implemented using fast incremental computations are available [5]. Figure 2 shows an example of an image segmented using an MDL-based algorithm.



(a) (b)

Figure 2: Parts of an object using an MDL-based segmentation algorithm.

Part Segmentation Using Compressed Images

In order to reduce time and memory requirements as well as to exploit frequency information made explicit by the Wavelet transform, we propose to extract parts directly from multiband semi-compressed images using the Wavelet transform. In [5] we described an approach to texture segmentation based on the MDL principle and the use of multiband Wavelet compressed images, developed by the PI during her summer research tour at Eglin AFB. The results obtained by the PI during this visit, show that this approach has the attractive properties that it is capable of robustly segmenting *textured* images while using *semi-compressed data.*

The main idea behind the algorithm is to optimize a cost function representing the length of encoding a segmentation of the given multiband image, into a set of non-overlapping regions that are homogeneous in a statistical sense. In particular, the algorithm encodes an image segmentation as a collection of regions modeled as polynomial surfaces of variable degree, perturbed by zero mean Gaussian noise and whose boundaries are described using a chain code representation. Thus, the

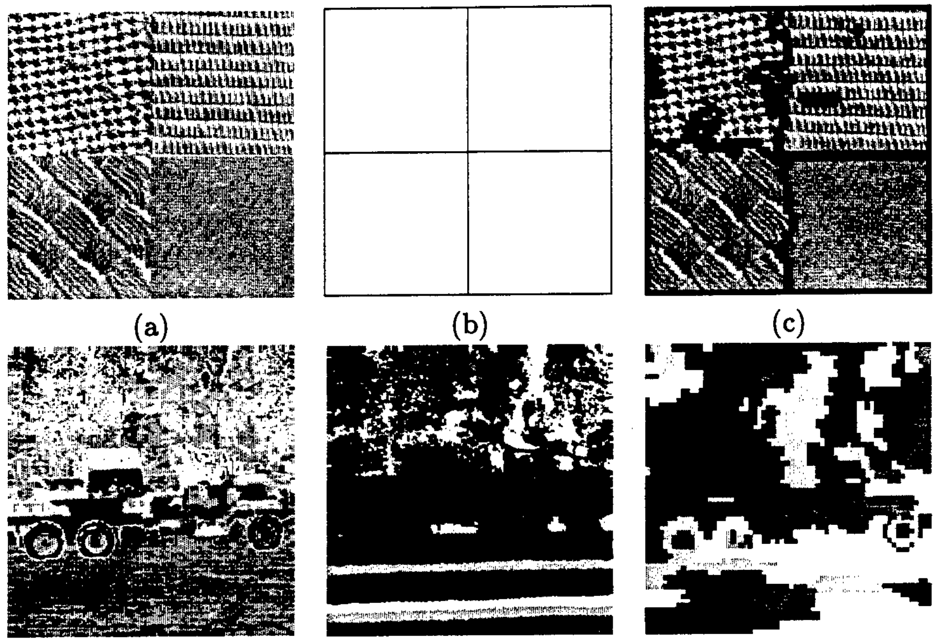


Figure 3: Examples. (a) Original 256×256 image composed of four textures. (b) Ideal segmentation of (a). (c) Segmentation of (a) obtained using a 64×64 resolution. (d) Original intensity ladar image. (e) Original range ladar image. (f) Segmentation of (d) and (e) using 8 64×64 intensity and range bands.

algorithm selects the image regions, as well as the degree and coefficients of polynomials that best fit the given data in each region. Since the algorithm uses the different bands of the image at the same time, it combines information about preferred orientations to distinguish between different textures.

Formally, consider an image with d bands, and let $\Omega = \{\omega_j\}$ denote the image segmentation into regions $\{\omega_j\}$ and let Y represent the image data and Y_j represent the image data within region ω_j. Further, assume that the image comes from a stochastic process that can be characterized as polynomial gray scale surfaces of unknown degree plus Gaussian noise described by a vector of parameters $\beta = \{\beta_j\}$. Then, the MDL objective function to optimize is given by:

$$L(Y, \Omega, \beta) = L(\Omega) + L(\beta|\Omega) + L(Y|\Omega, \beta). \tag{1}$$

where the first term is the length of encoding the region boundaries, the second term is the length of encoding the parameters and the last term is the length of encoding the residuals. If the boundaries are encoded using their chain code representation, assuming that at each point the number of possible directions is 3 (i.e. the number of adjacent grid points, excluding the current one), the first term of the encoding cost can be approximated by [31]:

$$L(\Omega) = \sum_i (l_i \log 3 + \log^*(l_i) + \log(2.865064))$$

where l_i is the length of the boundary i and $\log^*(x) = \log x + \log\log x + \log\log\log x + \ldots$ up to all positive terms. The second cost term, $L(\beta|\Omega)$, can be expressed using Rissanen's [31] expression

for optimal-precision analysis that says that K independent real-valued parameters characterizing n data points can be encoded using $(K/2)\log n$ bits. Thus,

$$L(\beta|\Omega) = \frac{1}{2}\sum_j K_{\beta_j} \log n_j$$

where K_{β_j} is the number of free parameters describing region j and is a function of the polynomial degree to be determined, and n_j is the number of pixels in region j. Finally, the third cost term $L(Y|\Omega,\beta)$ can be written using Shannon's theorems [1] as:

$$L(Y|\Omega,\beta) = -\log p(Y|\Omega,\beta) = \sum_j -\log p(Y_j|\beta_j)$$

The resulting objective cost function is in general non-convex and it is not possible to find a close solution to the optimization problem. In practice, a "good" local minimum is sought by searching in the segmentation space, using steepest decent from an initial segmentation. Thus, the segmentation algorithm must successively merge those pairs of neighbor regions that decrease the objective function the most. An initial segmentation could be, for example, the image itself, with each pixel considered a separate region. In this case, the initial number of regions that must be examined as candidates for merging, is equal to the number of pixels in the image. *The computational complexity of this search is reduced tremendously by using compressed data in a hierarchical structure.* This is achieved by starting the algorithm at a low resolution level, and using the resulting segmentation as the initial point at the next resolution level. Starting at a low resolution results in a reduction of the number of candidate regions by a factor of 4, 16, or more, depending on the number of levels in the hierarchy. Furthermore, propagating the resulting segmentations through the hierarchy levels allows the algorithm to start at each level with a relatively small number of regions.

Examples

Next, we show examples of segmentation using this algorithm. First, we show the results obtained segmenting an image artificially made by mosaicing four textures. This example is important, since the true segmentation is known by construction, and can be objectively compared with the obtained result. Figure 3(a) shows the original image, while Figure 3(b) shows the true segmentation, and Figure 3(c) shows the obtained segmentation using the proposed algorithm. It is seen that the four regions are correctly segmented except for a few small regions. The second example shows the segmentation of a ladar image of a real vehicle among heavy clutter. Figures 3(d) and (e) show the intensity and range images, respectively, and Figure 3(f) shows the result obtained segmenting 64×64 resolution images using the four bands of the intensity Wavelet compressed image and the four bands of the range Wavelet compressed image. It is seen that different parts of the target truck, such as its roof, hood, wheels, side door, cargo bed, etc. are all differentiated. Also the different textures of the background and floor are segmented.

Finally, Figures 4 and 5 show examples using color and range data captured with the new 3D scanner from Minolta purchased with funds from this grant. Figure 4 shows toy objects with curved surfaces, and Figure 5 shows scale models of cars. In both figures, the first row corresponds to the color image, the second row to the range image, the third row to the wavelet decomposition of the range data, and the last row to the obtained segmentation.

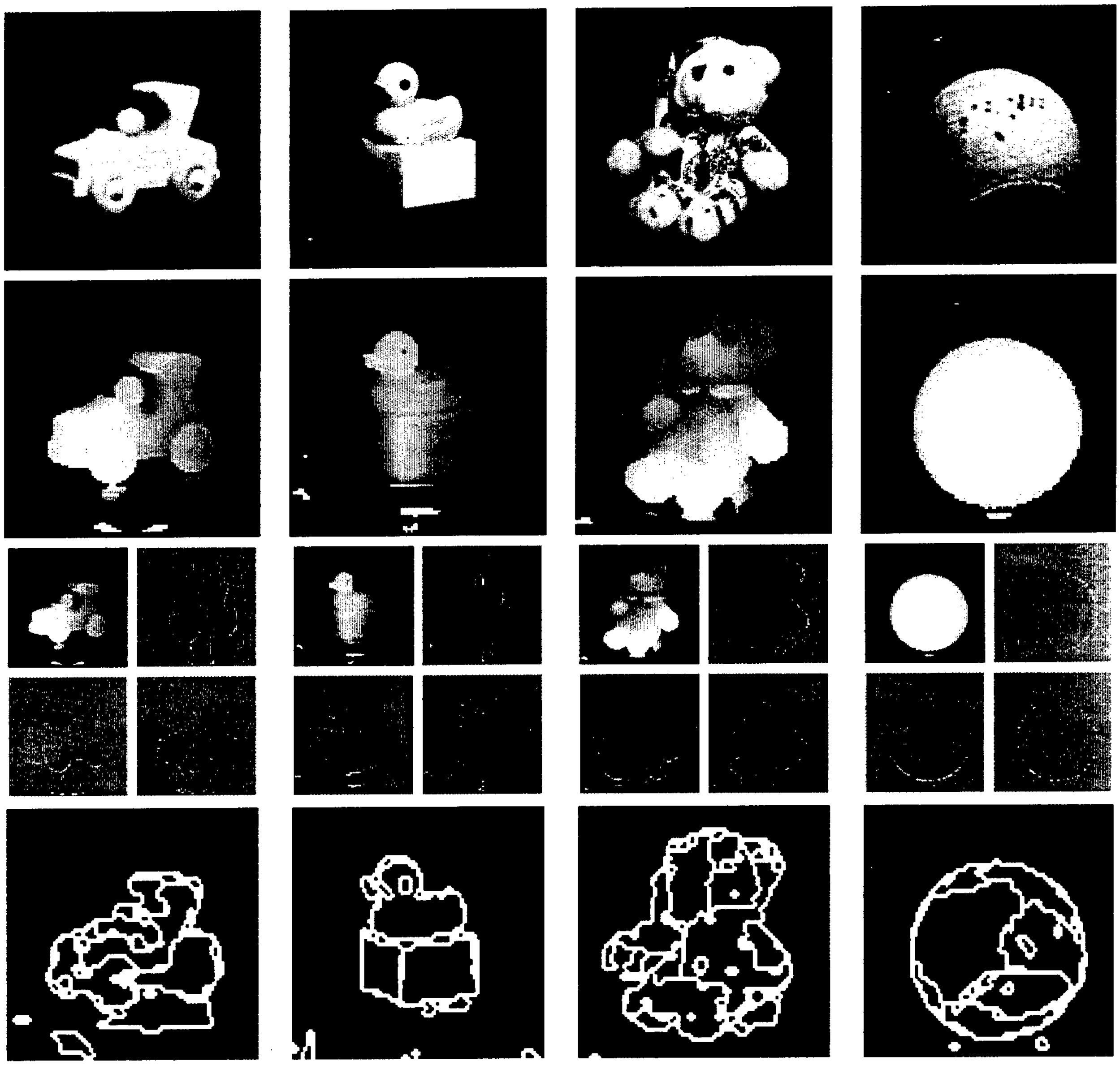


Figure 4: MDL segmentation of toys. (a) Color image; (b) Range image; (c) Wavelet decomposition of the range image; (d) MDL segmentation.

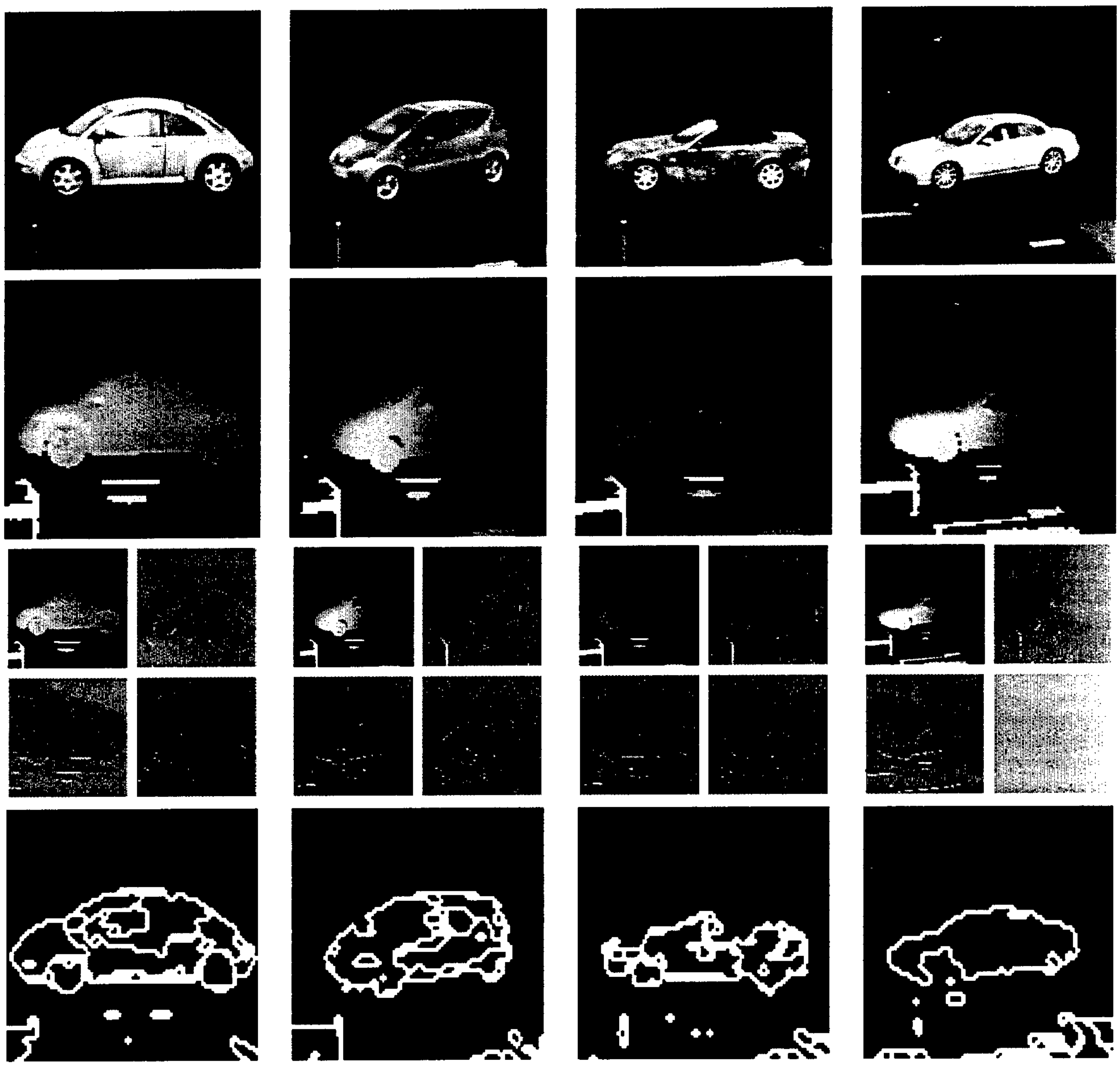


Figure 5: MDL segmentation of scale model of cars. (a) Color image; (b) Range image; (c) Wavelet decomposition of the range image; (d) MDL segmentation.

Appearances of Parts and their Relations

Obviously, parts obtained using the definition given above are sensor and illumination dependent. Thus, in order to completely characterize an object for different sensors and light sources, we introduce the concept of "appearances" of a part. Two parts segmented from two images of the same object, obtained with similar sensor and illumination configurations, are said to be **appearances** of the same part if they are statistically similar and are in similar image locations.

The effects of the sensor and illumination configurations on the appearance of a part are learned by tracking the part in sequences of compressed images spanning the space of all possible configurations. The statistical similarity between parts from consecutive frames is measured from the statistics provided by the segmentation algorithm discussed above. This concept can be formalized as follows. Let ω_1 and ω_2 be two parts obtained from two images with d bands of the same object with different, but similar, sensor and illumination configurations. Let Y_1 be an $n_1 \times d$ matrix with the intensity pixel values in part ω_1 and Y_2 be an $n_2 \times d$ matrix with the intensity pixel values of part ω_2. Let q be the order of the polynomials used to fit the parts, and $m = (q+1)(q+2)/2$ be the number of polynomial coefficients. Let Φ_1 and Φ_2 be an $n_1 \times m$ and an $n_2 \times m$ matrices of m basis functions spanning the polynomials, evaluated at each of the n_1 and n_2 pixels – i.e. products of powers of pixel row and column coordinates – respectively. Finally, let Θ_1 and Θ_2 be two $m \times d$ matrices with the *optimal* regression coefficients for ω_1 and ω_2, respectively. Using these definitions, we have [16]:

$$Y_i = \Phi_i \Theta_i + \Psi_i \;\; i = 1, 2$$

where Ψ_1 and Ψ_2 are vectors of zero mean Gaussian noise with covariance $\sigma^2 I$, and Θ_1 and Θ_2 are estimated by minimizing the expected fitting error:

$$\epsilon_i = \|Y_i - \Phi_i \Theta_i\| \;\; i = 1, 2$$

Then, the two parts ω_1 and ω_2 are considered appearances of the same part ω if

$$\epsilon_{1,2} = \frac{1}{n_1}\|Y_1 - \Phi_1 \Theta_2\| + \frac{1}{n_2}\|Y_2 - \Phi_2 \Theta_1\| \leq T_\epsilon$$

and

$$\Delta_{1,2} = \|\mu_1 - \mu_2\| \leq T_\Delta$$

where μ_1 and μ_2 are the centroids of the parts and T_ϵ and T_Δ are given thresholds. As part of this research, we are studying how to automatically set these thresholds based on the noise model and knowledge of the sensor locations. Note that the proposed criteria is able to handle over and under segmentation problems by assigning more than one part in one frame to a part in the other frame. An alternative to the above criteria that we are exploring is to use the MDL principle between frames to decide which regions are appearances of the same part.

Appearances as manifolds

The collections of these appearances need to be stored compactly and efficiently retrieved at recognition time. For this, we construct parametrized manifolds interpolating the projections of the individual appearances into eigenspaces obtained by applying the Karhunen-Loeve compression method [27] to a scale and brightness normalized set of the appearances. These manifolds are

similar to the ones proposed in [26], which have been shown to be very successful when used to recognize and locate single objects. However, until now they have been used to represent appearances of complete objects and therefore they have failed in the presence of occlusion. As noted in section 3 recent attempts to overcome occlusion [23, 22] have been not successful when there is also object translation and scaling. In this research, we use this type of representation with *parts*, to take advantage of their good localization properties while addressing the occlusion problem *even in the presence of object scaling and translation.*

Let $p_1, p_2, \ldots, p_n$ be the set of collected appearances of parts, where p_i is an $N \times 1$ vector with the pixel values of the image of an appearance. Let S be the $N \times N$ covariance matrix of the training appearances:

$$S = QQ^T$$

where

$$Q = [\ \hat{p}_1 \hat{p}_2 \cdots \hat{p}_n\ ] \tag{2}$$

and

$$\hat{p}_i = p_i - \frac{1}{n}\sum_{j=1}^{n} p_j$$

Then, an appearance can be expressed as a linear combination of $M << N$ eigenvectors of the covariance matrix S:

$$\hat{p} \sim \tilde{p} = \sum_{i=1}^{M} a_i e_i$$

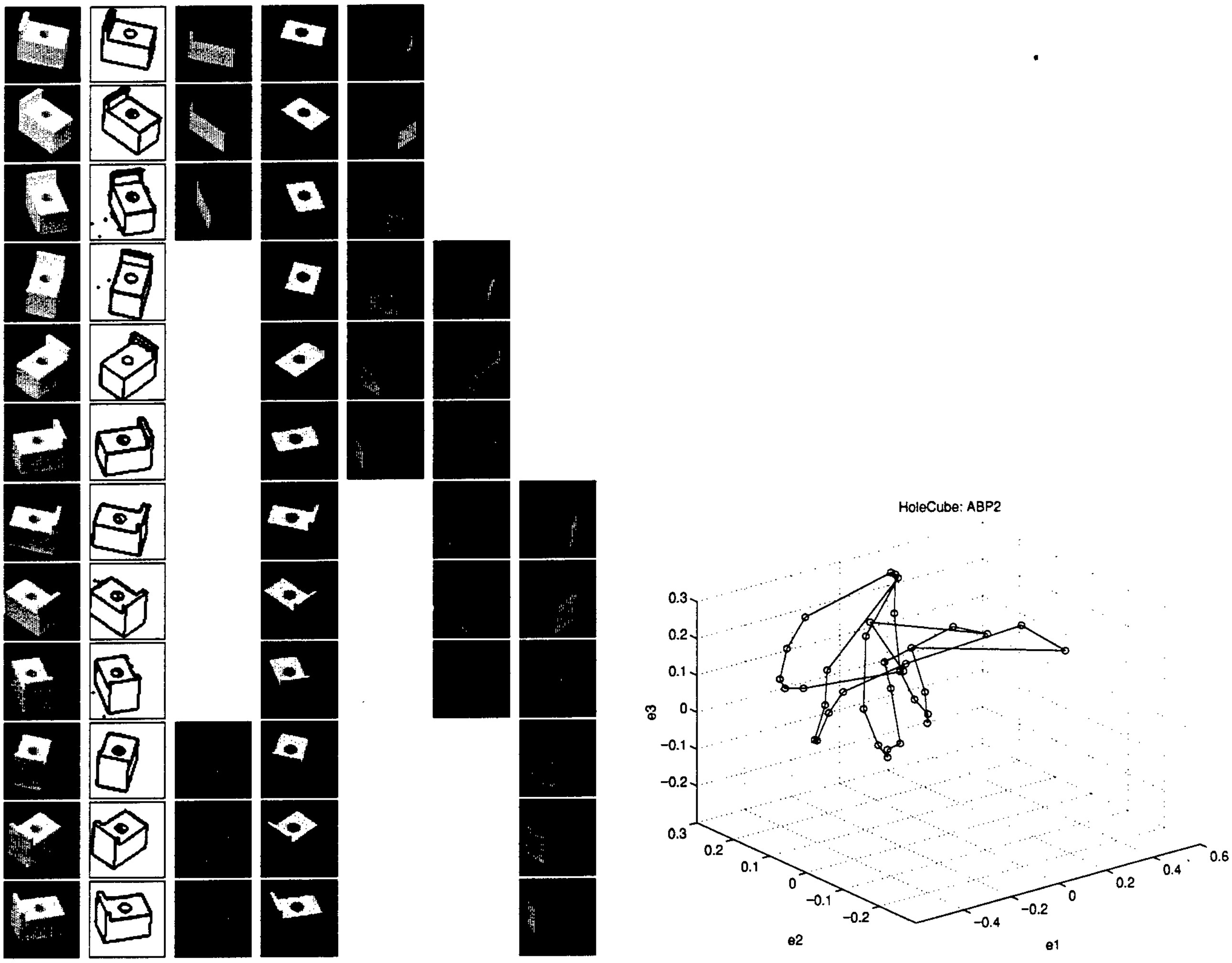
where e_i, $i = 1, 2, \ldots M$, are the $N \times 1$ eigenvectors of S with the M largest eigenvalues, and a_i, $i = 1, 2, \ldots M$, are scalar coefficients. Thus, as different appearances of a part are generated by varying viewpoint or illumination conditions, their Karhunen-Loeve reductions sweep a manifold in an M-dimensional space.

Figure 6 illustrates the appearances of parts of the object "HoleCube". Figure 6(a) shows images of this object every 30^o and Figure 6(b) shows their respective MDL segmentations. Figure 6(c) shows the appearances of five of its parts. Note that due to self-occlusion, four of the parts disappear for some frames. Figure 6(d) shows the manifold corresponding to the second part of "HoleCube", visualized in three dimensions.

Using this representation, part detection can be accomplished by taking a region segmented from the image and projecting it into the eigenspace spanned by the first M eigenvectors of S. If the distance between the given region and its projection is sufficiently small, the part is said to be present and its parameters are obtained by searching the closest training image to the projection.

Relations of ABPs

Although it is possible to identify some objects by recognizing some of their distinctive ABPs, recognizing general objects having several "common" parts requires the use of spatial relationships between the parts being recognized. Furthermore, in the context of recognition in the presence of clutter and occlusion, spatial relationships can play a very important role. We represent spatial relationships between ABPs, *appearance-based relations* (ABRs) using eigenspaces similar to the



(a) (b) (c) (d)

Figure 6: Collection of appearances of parts for "HoleCube". (a) Images of "HoleCube" every 30^o. (b) MDL segmentations of the images in (a). (c) Appearances of five parts. (d) Manifold corresponding to the second part, shown as three dimensional for visualization purposes.

ones used to represent ABPs. ABRs are developed by merging adjacent ABPs to create new training sets that are also represented as manifolds in the corresponding spanned eigenspace.

ABP and ABR Representation Using Compressed Images

The proposed representation using eigenspaces can be used *directly on images compressed with the Wavelet transform*. This can be easily shown by realizing that the distances in the ABP and ABR eigenspaces are preserved under linear transformations and image subsampling. Consider the K-L reduction $\tilde{p}$ of the normalized appearance $\hat{p}$:

$$\tilde{p} = \sum_{i=0}^{M} a_i e_i$$

and a linear filter F such as the low pass or high pass filters used to generate the Wavelet decomposition of $\hat{p}$. Then, from linearity we have

$$\tilde{p} * F = \sum_{i=0}^{M} a_i (e_i * F) \tag{3}$$

Furthermore, equation 3 is valid pointwise, at every pixel of the images involved. Thus, the equality is not affected by down sampling:

$$(\tilde{p} * F)_{\downarrow} = \sum_{i=0}^{M} a_i (e_i * F)_{\downarrow}$$

Therefore, the K-L coefficients a_i, $i = 1, \ldots, M$, are the same for the original appearance and for its filtered and down sampled versions, such as its Wavelet decomposition components. Moreover, the eigenvectors used to K-L reduce the Wavelet components of the image are the Wavelet components of the eigenvectors used to K-L reduce the original images.

Quantization effects due to compression, are bounded and well behaved. For small perturbations of the covariance matrix S, provided that its eigenvalues are well separated, its eigenvectors change little. Consider a perturbation of the training matrix Q defined in equation 2, $\tilde{Q}$:

$$\tilde{Q} = Q + \Delta$$

where Δ represents a perturbation introduced by quantization errors. Then, the new covariance matrix is given by

$$\tilde{S} = \tilde{Q}\tilde{Q}^T = S + E$$

where

$$E = \Delta Q^T + \Delta\Delta^T + Q\Delta^T$$

Since both S and E are symmetric, applying the Wielandt-Hoffman theorem [11], the perturbation of the eigenvalues of S is bounded by

$$\sum_{i=1}^{N} (\lambda_i(\tilde{S}) - \lambda_i(S)) \leq \|S\|_F^2 \tag{4}$$

where $\|S\|_F^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij}^2$. On the other hand, the perturbation on the eigenvectors of S depends not only on the perturbation E but also on the separation between the eigenvalues of S. Let G be a matrix such that $E = \epsilon G$ and $\|G\|_2 = 1$. Then, the eigenvector perturbation is given by [11]:

$$\tilde{e}_k - e_k = \epsilon \sum_{i=1, i \neq k}^{N} \frac{x_i^T G x_k}{\lambda_k - \lambda_i} x_i + O(\epsilon^2) \tag{5}$$

In the case that there are repeated eigenvalues, the perturbation can be bound using subspaces, instead.

From the above, we conclude that the ABP and ABR representations can be stored in compressed form and that recognition can be robustly performed without decompressing the given image nor the eigenvectors of the representation. Moreover, equations 4 and 5 make explicit the tradeoff between quantization rate (compression rate, data loss) and classification performance.

4.2 Mathematical models for robust recognition

We believe that the proposed ABPs and ABRs are rich features with many properties that can be used to overcome problems with data uncertainty and occlusion. In the second stage of the proposed research we are studying them and their properties to develop mathematical models for the design of robust object recognition systems, such that the probabilities of false alarms and misdetection are minimized. In particular, we will model from first principles and appropriate noise models the discriminatory power of ABPs and ABRs. These mathematical models, which will be statistically validated through experiments with real and simulated data, will be used in a Bayesian framework to organize large object databases and to achieve robust recognition in the presence of data uncertainty.

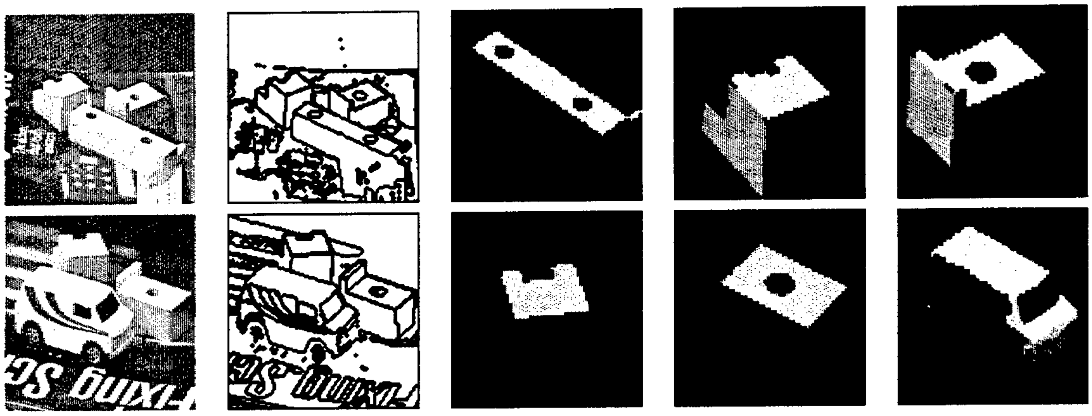
Generation of ABP and ABR identity and pose hypotheses

An image is an *observation* of a subset of the models. Not all the ABPs and ABRs in the models database participate in the observation. Furthermore, not all the regions and pairs of regions in the image are related to the ABPs and ABRs in the database. Let $\mathcal{ABP}$ and $\mathcal{ABR}$ be the sets of the union of the ABPs and ABRs manifolds, respectively, for *all* the objects in a given database. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the set of the projections of the parts or regions of a given image into the ABP eigenspace, and the set of the projections of pairs of adjacent parts into the ABR eigenspace, respectively. The *recognition* problem is to find two unknown correspondence mappings $h_{ABP} : \mathcal{S}_1 \rightarrow \mathcal{ABP}$ and $h_{ABR} : \mathcal{S}_2 \rightarrow \mathcal{ABR}$ associating ABPs and ABRs to image regions and pairs of regions, respectively. The *localization* problem is to find two unknown correspondence mappings $l_{ABP} : \mathcal{S}_1 \rightarrow \mathcal{R}_1^m$ and $l_{ABR} : \mathcal{S}_2 \rightarrow \mathcal{R}_2^m$ associating *appearances* of ABPs and ABRs to image regions and pairs of image regions, respectively.

Given an image and its MDL segmentation, the identity correspondence mappings h_{ABP} and h_{ABR} can be generated by projecting each segmented region and pair of regions into the eigenspaces obtained during training, and finding points on manifolds near to these projections. While the selected manifolds provide hypotheses for the part and relation identities, the selected points on these manifolds provide the localization mappings l_{ABP} and l_{ABR} with the pose hypotheses. Furthermore, the actual distances between the projections and the manifolds and points are quantitative measures that can be used to estimate the reliability of the hypotheses as well as to combine them using a Bayesian framework.

Preliminary Results

Preliminary results are very encouraging. We have explored the potential of ABPs and ABRs by implementing a very simple minded recognition system that given a part p, takes those ABP hypotheses with distance $d = d(p, h_{ABP}(p)) \leq T_1$, where T_1 is a "small" threshold as successful hypotheses. Other ABP hypotheses with somewhat larger distances $T_1 \leq d \leq T_2$, where T_2 is a second threshold are verified or discarded by using ABRs hypotheses. An ABR hypothesis for a pair of image parts $r = (p_1, p_2)$ is said to verify the ABP hypotheses for the component parts p_1 and p_2 if the distance $d(r, h_{ABR}(r)) \leq T_3$, where T_3 is a third threshold.



(a) (b) (c)

Figure 7: Results for cluttered scenes. (a) Cluttered scenes. (b) MDL segmentations. (c) ABP and ABR hypotheses.

Two examples with background clutter are shown in Figure 7. The first column shows the original images, the second column shows the corresponding MDL segmentations and the following columns show the appearances of the ABPs and ABRs that were hypothesized and verified, using this simple algorithm. It can be seen that in spite of the clutter, partial occlusion of the objects, and segmentation problems such as the merging of some of the object parts with the background, all the objects and their pose were correctly identified. In this preliminary study, the thresholds T_1, T_2 and T_3 were arbitrarily set. A study of the effects of these thresholds has also been done by varying them while processing 12 images of each of the above scenes, taken in increments of 30 degrees angles. The results of these experiments showed that the best performance was achieved when $T_1 = 0.03$, $T_2 = 0.05$ and $T_3 = 0.08$ with probabilities of false alarm and misdetection of the order of 0.2. However, a part of the proposed research is to design probabilistic models that will enable us to eliminate the use and tuning of *ad hoc* thresholds such as the ones used in this example, as discussed below.

Discriminant Power of ABPs and ABRs

A possible choice for the correspondence mappings is to assign to each image region and pair of regions the closest manifold in the database to its projection, and the closest point on this manifold. However, it is possible for more than one manifold to be close enough to the projection of the given image region or pair of regions, making the choice of the nearest manifold somehow arbitrary. A more reasonable approach is to consider more than one possible assignment, ranked in terms of their probability of being correct. This observation leads us to believe that a recognition system should use information about the ABPs and ABRs discriminant power, where the discriminant power is measured in terms of how far the considered manifold is to other manifolds of the same type.

We are modeling the ABPs and ABRs as being the result of a stochastic process, where each manifold can be expressed as a nominal manifold plus noise. This requires to model the noise

component, to estimate its parameters from sample data, and to statistically validate it. The basic principles that we are following are outlined next. These results apply to both ABPs and ABRs, and hence we talk of manifolds and appearances without referring to their particular type.

Let M_k denote an instance of the k^{th} manifold of a given type[1] and a_{kj} denote its j^{th} appearance. The probability of misclassifying appearances of M_k as appearances of M_i is given by

$$P(M_i|M_k) = \frac{\sum_j \sum_m P(a_{im}|a_{kj})P(a_{kj})}{P(M_k)}$$

This probability should be large if the manifolds M_i and M_k are close to each other for most of the appearances of M_k. Assuming that all appearances of a given type are equally likely we have:

$$P(a_{kj}) = \frac{1}{\#\text{ appearances in the database}} \quad \text{and} \quad P(M_k) = \frac{\#\text{ appearances of } M_k}{\#\text{ appearances in the database}}$$

On the other hand, we are currently exploring how to model the probability $P(a_{im}|a_{kj})$. For this, we are pursueing two approaches in parallel. On one hand, we will propagate analytically simple data noise models, such as Gaussian models, through the eigenspace reduction and the Wavelet compression process. On the other hand, we will use non-parametric methods such as bootstrapping to estimate the noise from experimental data. Statistical model validation will be performed using resampling methods [12].

The discriminant power index between manifolds M_i and M_k can then be defined as

$$DPI(M_i|M_k) = 1 - P(M_i|M_k)$$

In this way, the discriminant power index of M_k from M_i is large whenever the probability of misclassifying M_k as M_i is low.

As a preliminary indicator of how this index may work we approximated the probability $P(a_{im}|a_{kj})$ as a normalized decaying exponential function of the distance between the corresponding projections. Using this approximation, the discriminant power indices of the manifolds ABP3 and ABP5 of "HoleCube" in the eigenspace spanned by its ABPs are 0.125 and 0.059, respectively. These low values are not surprising, since these manifolds represent identical rectangular faces in "HoleCube" and therefore have very similar appearances, making it very difficult to discriminate between them – i.e. they have very low relative discriminant power.

Organization of Large Databases of Objects

A major concern with the new representation is whether or not it will scale up to be used with a large database of objects, since each object can potentially have many ABPs and ABRs. We believe that this problem can be addressed by using some of the results from the previous discussion. Specifically, we propose to use a hierarchical organization scheme of ABPs. In this hierarchy, ABPs will be organized in a tree structure, as follows. At the top level, all ABPs will be grouped together and they will span the *Universal* eigenspace; then, at each node in the tree the ABPs will be divided into groups, each of them consisting of more similar (with lower discriminant power indices) parts.

[1] A manifold and an appearance can be either of ABP or ABR type.

In turn, each of these groups will define a new eigensubspace where the highest order coordinates will provide more discriminatory power between the ABPs spanning the space.

This organization will allow different objects to "share" similar ABPs with the consequent reduction in storage requirements while also reducing the matching complexity. Furthermore, this structure provides a natural way to incrementally expand the database of ABPs. Whenever a new ABP needs to be added it can be compared with the existing groups to decide to which group it should belong or to create a new group if necessary. A similar hierarchy will be maintained for the appearance based relations.

4.3 Experimental Protocol and Performance Characterization

A scene is affected by the pose and illumination of the objects. Occlusion, imperfect segmentations and interobject reflections are important factors that will influence the performance of the system. Imperfect segmentations might happen due to occlusion and the change of pose. Interobject reflections will occur when a surface reflects other surface in its environment. Also, as explained in section 4.1 there is an unavoidable tradeoff between compression rate and system performance.

As a part of this research, we will develop a rigorous experimental protocol using synthetic and real data to characterize the performance of the system in terms of scene complexity (measured in number of objects), amount of clutter (measured in number of unrelated objects), compression rate, and probabilities of misdetection, misclassification and pose error.

The most common tool used to characterize the performance of a detection algorithm is a plot of its probability of misdetection versus its probability of false alarm, as some tuning parameter is changed. This plot is commonly known as the "receiver operating curve" of the system, or ROC, for short. Although ROCs are useful by representing the system performance as a parameter is varied, they have several limitations. One disadvantage in using ROCs is due to the fact that only one parameter can be varied at a time. Thus, if the effect of variations of multiple variables needs to be studied, a different curve must be determined for each of these variables making the analysis of the system performance more difficult. A second disadvantage is that it is difficult to compare ROCs for different algorithms since they may take different variables into account. Finally, obtaining ROCs is an expensive process where factorial experiments must be carried out to determine the system performance at all performance levels from probability of false alarm equal to 0 to probability of false alarm equal to 1. Thus, we propose to characterize the performance of the target detecting algorithms by using a methodology similar to the one described in [20]. This methodology, which was adapted from the psychology literature and is discussed next, provides an alternative characterization tool to summarize multiple ROCs into a single curve, solving the problems described above.

Consider a detection algorithm that must report whether a given image has a target or not. Typically, the algorithm would compute some measure of evidence of target presence and compare it to some given threshold value. Whenever the evidence measure is greater than the given threshold, a target would be reported. The performance of the algorithm is affected by several factors as descirbed above. The effect of variations of these variables on the overall performance can be measured through the use of equivalent effects of some critical signal variable by following the four

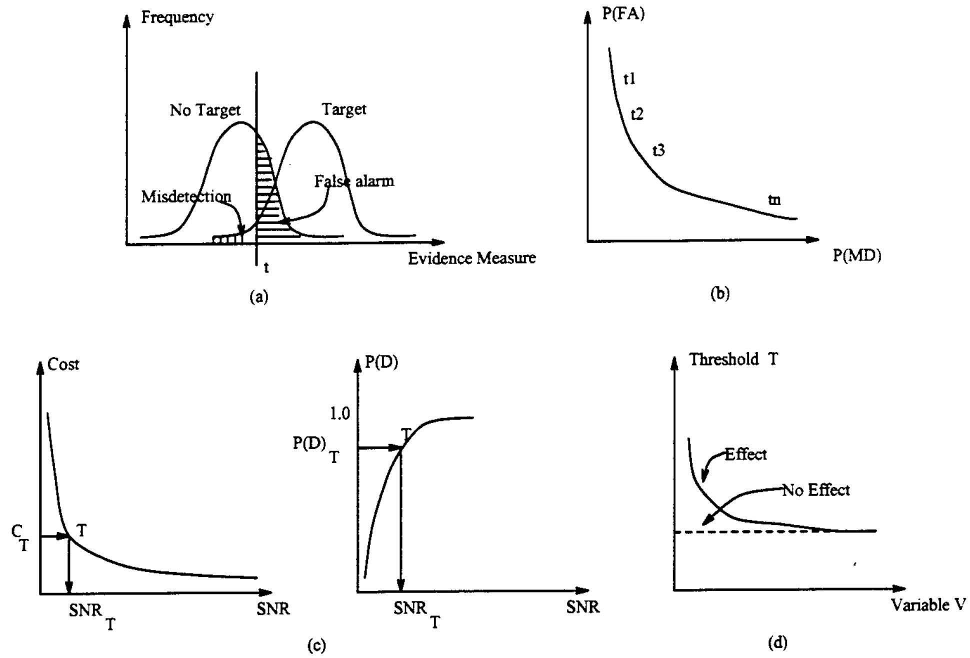


Figure 8: Steps for performance characterization. (a)Step 1: Obtain the frequency distributions of the evidence measure for images with target and no target. (b) Step 2: Obtain the ROC. (c) Step 3: Determine the optimal functioning point using either the expected cost or the probability of detection given the probability of false alarm. (d) Step 4: Plot the threshold value corresponding to the optimal operating point versus a variable of interest.

steps described below.

1. **Obtain evidence distributions.** The first step consists on obtaining two distributions of evidence measures, one for images with target and another for images without target, as illustrated in Figure 8(a). This can be done by randomly presenting the algorithm with images of both types and recording the frequency of the evidence measure values reported by the algorithm.

2. **Obtain ROCs.** The second step consists on constructing an ROC as the one shown in Figure 8(b) by varying the threshold used by the algorithm to compare against the computed evidence measure. False alarms occur when the given image does not contain a target, but the evidence measure is greater than the threshold being used. Misdetections occur when the given image contains a target, but the evidence measure is less than the threshold. The probabilities of false alarms and misdetections can be approximated by their frequency ratios:

$$
\begin{aligned}
P(\textit{false alarm}) &= P(\textit{target}|\textit{no target}) = \frac{\text{Number of false alarms}}{\text{Total number of input images with no target}} \\
P(\textit{misdetection}) &= P(\textit{no target}|\textit{target}) = \frac{\text{Number of misdetections}}{\text{Total number of input images with target}}
\end{aligned}
$$

3. **Determining the optimal operating point.** The optimal operating point (or its corresponding threshold value) can be specified in different ways, depending on how much prior knowledge is available. If the prior probabilities of target and no target and the costs of false

alarms, misdetection and detection are known, the optimal operating point can be defined as the one minimizing the expected cost. Let C_{tn}, C_{nt}, C_{tt}, and C_{nn}, be the cost of a false alarm, the cost of a misdetection, the cost of correctly detecting a target, and the cost of correctly rejecting a target, respectively. Then, the expected cost is given by:

$$\begin{aligned} E(C) &= \left[P(\text{target}|\text{no target})C_{tn} + P(\text{no target}|\text{no target})C_{nn}\right] P(\text{no target}) + \\ &\quad \left[P(\text{target}|\text{target})C_{tt} + P(\text{no target}|\text{target})C_{nt}\right] P(\text{target}) \end{aligned}$$

and the optimal operating point is found by minimizing $E(C)$ where the minimization variable is the threshold to be used by the algorithm. In the most likely case when the costs are difficult to set, an alternative way to define the optimal operating point is to use the Neyman-Pearson criterion – i.e to maximize the probability of detection for a *given* probability of false alarm.

Independently of which definition is used, the optimal operating point depends on the signal to noise ratio (SNR) in the input image. For example, increasing the image contrast results on an increase of the SNR and, hopefully, in an improvement of the algorithm performance for a given threshold value. The optimal operating points for different SNRs can be found by repeating steps 1 and 2 for the corresponding SNR values and determining the optimal point for each of the resulting ROCs. Once this is done, a graph of the expected cost or the probability of detection versus SNR can be plotted, depending on which definition of operating point is being used. This is illustrated in Figure 8(c). Finally, let SNR_T and T be the SNR and the associated threshold values for the optimal operating point for a given level of performance, as shown in the figure. The level of performance is specified by either a desired expected cost of classification or a desired probability of misdetection, again, depending on which optimal criterion is used.

4. **Performance analysis with respect to variables of interest.** Besides SNR, other factors affect the algorithm performance and merit study. Examples are the compression rate, the size of the target and the amount and nature of image clutter. In order to study these effects, steps 1 to 3 are repeated for different values of variables representing these variations. These results are then summarized in a graph where the threshold T determined in step 3 is plotted against the value of the variable of interest, as shown in Figure 8(d). A fairly flat plot indicates that the effect of the variable being considered on the optimal operating point of the algorithm is negligible. On the other hand, a steep plot indicates that the variable has a high impact on the performance.

5 Impact of the Research

The rapid changes that have taken place in our world over the past few years have significantly changed the roles and the requirements of the U. S. Air Force. The well-defined global threat posed by the Soviet Union's conventional and nuclear forces has been replaced by numerous, localized, potential threats which may involve biological and chemical weapons as well as conventional and nuclear ones. In addition, the Air Force routinely engages in defense operations other than warfare including peacekeeping and humanitarian aid roles. Other enormous changes have taken place in our society as well. Information technology is revolutionizing the way we live our lives. Intelligence, counter-intelligence and target acquisition (TA) are the fundamental components of warfare in the information age.

Imaging sensors and image analysis systems will play an increasingly important role as the Air Force's use of automated, unmanned ground and aerial vehicles and robots for high-risk operations increases. Use of these vehicles for operations other than warfare will require significantly higher levels of reliability and functionality than is currently available. Advanced, realistic, and comprehensive scene modeling, and robust, reliable methods for automatic target/object detection, recognition, classification and identification are key technology areas for meeting the Air Force's requirements.

Funds from this grant have provided the needed infrastructure to achieve these goals. Once completed, this research effort will result in a object recognition system capable of recognizing free-form, possibly articulated, 3D objects in semi-compressed images of cluttered scenes and whose performance will be rigorously characterized. The major thrust of this research takes advantage of the PI experience, her previous results on CAD-based vision, and the results obtained by the PI during her summer research visit to Eglin AFB. The main contribution of the approach is an automatic object recognition system that:

- works well in the presence of clutter and occlusion,
- minimizes the probabilities of false alarms and misdetections,
- does not use *ad hoc* parameters,
- is computationally efficient since it works directly on compressed data,
- can be incrementally trained,
- is suitable for different imaging modalities such intensity and/or ladar images, and
- has its performance rigorously characterized using extensive testing.

6 Dissemination of the results

We are currently assembling a large database with registered color and range data of hundreds of objects. This data will be shortly made available to the research community through the world wide web. Furthermore, we are in close collaboration with The Ohio State University computer vision research group, who also owns a Minolta 3D scanner, to join our databases.

The new object representation has been reported in the article:

"Object Representation Using Appearance-Based Parts and Relations," O. I. Camps, C. Y. Huang, N. Pande and T. Kanungo, *submitted* to *IEEE* Trans. Pattern Analysis and Machine Intelligence.

References

[1] N. Abramson. *Information Theory and Coding.* McGraw Hill, 1963.

[2] F. Arman and J. K. Aggarwal. CAD-based object recognition in range images using pre-compiled strategy trees. In A. Jain and P. Flynn, editors, *Three-Dimensional Object Recognition Systems*, pages 115–134. Elsevier Science Publishers, 1993.

[3] T. O. Binford. Body-centered representation and perception. In *Lecture Notes in Computer Science (994): Object Representation in Computer Vision.* Springer-Verlag, 1995.

[4] O. I. Camps. Towards a robust physics based object recognition system. In *Lecture Notes in Computer Science (994): Object Representation in Computer Vision*, pages 297–312. Springer-Verlag, 1995.

[5] O. I. Camps. MDL texture segmentation of compressed images. *Final Report for Summer Faculty Research Program, Wright Laboratory*, pages 9–1–9–20, 1997.

[6] O. I. Camps, L. G. Shapiro, and R. M. Haralick. Image prediction for computer vision. In Jain A.K. and P.J. Flynn, editors, *Three-dimensional Object Recognition Systems.* Elsevier Science Publishers BV, 1993.

[7] O. I. Camps, L. G. Shapiro, and R. M. Haralick. A probabilistic matching algorithm for computer vision. *Annals of Mathematics and Artificial Intelligence*, 10:85–124, 1994.

[8] M. S. Costa and L. G. Shapiro. Scene analysis using Appearance-Based Models and Relational Indexing. In *International Symposium on Computer Vision*, pages 103–108, Florida, November 1995.

[9] S. J. Dickinson. Part-based modeling and qualitative recognition. In A. Jain and P. Flynn, editors, *Three-Dimensional Object Recognition Systems*, pages 201–228. Elsevier Science Publishers, 1993.

[10] V. T. Franques and D. A. Kerr. Wavelet-based rotationally invariant target classification. In *SPIE Conference*, 1997.

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins Univesity Press, Baltimore, 1989.

[12] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer-Verlag, New York, 1994.

[13] W. E. L. Grimson, T. L. P., S. J. White, and N. Noble. Recognizing 3d objects using constrained search. In A. Jain and P. Flynn, editors, *Three-Dimensional Object Recognition Systems*, pages 259–284. Elsevier Science Publishers, 1993.

[14] M. Hebert, J. Ponce, T. Boult, and A. Gross. Report on the 1995 Workshop on 3-D Object Representations in Computer Vision. In *Object Representation in Computer Vision- Lecture Notes in Computer Science 994*, pages 1–18. Springer-Verlag, 1995.

[15] M. Hebert, J. Ponce, T. Boult, and A. Gross. Report on the 1995 workshop on 3d object representations in computer vision. In *Lecture Notes in Computer Science (994): Object Representation in Computer Vision.* Springer-Verlag, 1995.

[16] C. Y. Huang, O. I. Camps, and T. Kanungo. Object Recognition Using Appearance-Based Parts and Relations. In *Proc. IEEE Computer Vision and Pattern Recognition*, Puerto Rico, June 1997.

[17] D. Huttenlocher. Recognition by alignment. In A. Jain and P. Flynn, editors, *Three-Dimensional Object Recognition Systems*, pages 311–326. Elsevier Science Publishers, 1993.

[18] D. Huttenlocher. Using two-dimensional models to interact with the three dimensional world. In *Lecture Notes in Computer Science (994): Object Representation in Computer Vision.* Springer-Verlag, 1995.

[19] A. K. Jain and P. J. Flynn. *Three-Dimensional Object Recognition Systems.* Elsevier, 1993.

[20] T. Kanungo, M. Y. Jaisimha, J. Palmer, and R. Haralick. A methodology for quantitative performance evaluation of detection algorithms. *IEEE Trans. on Image Processing*, 4(12):1667–1674, 1995.

[21] D. Kriegman and J. Ponce. Representations for recognizing complex curved 3d objects. In *Lecture Notes in Computer Science (994): Object Representation in Computer Vision.* Springer-Verlag, 1995.

[22] J. Krumm. Eigenfeatures for planar pose measurement of partially occluded objects. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 55–60, San Francisco, California, June 1996.

[23] A. Lenardis and H. Bischof. Dealing with occlusions in the eigenspace approach. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 453–458, San Francisco, California, June 1996.

[24] D. Metaxas. A physics-based framework for segmentation, shape and motion estimation. In *Lecture Notes in Computer Science (994): Object Representation in Computer Vision.* Springer-Verlag, 1995.

[25] J. Mundy, A. Liu, N. Pillow, A. Zisserman, S. Abdallah, S. Utcke, S. Nayar, and C. Rothwell. An experimental comparison of appearance and geometric model based recognition. In *Lecture Notes in Computer Science (1144): Object Representation in Computer Vision II.* Springer-Verlag, 1996.

[26] H. Murase and S. K. Nayar. Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, January 1995.

[27] E. Oja. *Subspace methods of Pattern Recognition.* Research Studies Press, Hertfordshire, 1983.

[28] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(7):715–729, July 1991.

[29] J. Ponce, A. Zisserman, and M. Hebert. Report on the 1996 international workshop on object representation in computer vision. In *Lecture Notes in Computer Science (1144): Object Representation in Computer Vision II.* Springer-Verlag, 1996.

[30] A. R. Pope and D. G. Lowe. Learning appearance models for object recognition. In *Lecture Notes in Computer Science (1144): Object Representation in Computer Vision II.* Springer-Verlag, 1996.

[31] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):211–222, 1983.

[32] C. A. Rothwell. *Object Recognition through Invariant Indexing.* OUP, 1995.

[33] S. Sclaroff and A. P. Pentland. Modal Matching for Correspondence and Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(6):545–561, June 1995.

[34] M. Zerroug and G. Medioni. The challenge of generic object representation. In *Lecture Notes in Computer Science (994): Object Representation in Computer Vision.* Springer-Verlag, 1995.

[35] A. Zisserman, D. Forsyth, J. Mundy, C. Rothwell, J. Liu, and N. Pillow. 3D object recognition using invariance. *AI Journal*, (78):239–288, 1995.